UNITED STATES DISTRICT COURT
DISTRICT OF MASSACHUSETTS

SCANSOFT, INC., Plaintiff v. VOICE SIGNAL TECHNOLOGIES, INC., LAURENCE S. GILLICK, ROBERT S. ROTH, JONATHAN P. YAMRON, and MANFRED G. GRABHERR, Defendants	C.A. No. 04-10353-PBS

DECLARATION OF BRUCE BALENTINE
CONCERNING CLAIM CONSTRUCTION OF U.S. PATENT NO. 6,594,630

I, Bruce Balentine, declare as follows.

1. I am a consultant in the speech recognition industry, in which I have worked for the past twenty years. I have designed and developed many commercial applications of speech recognition systems for Fortune 500 companies. I am also the inventor or co-inventor of several patents in this field. I have authored a best-selling book and other publications on the design of speech recognition applications and have lectured widely on this subject at speech recognition industry workshops, conferences, seminars, and trade shows. I am currently Chief Scientist and Executive Vice President of Enterprise Integration Group ("EIG"), a consulting, research, and engineering firm specializing in the design of Interactive Voice Response ("IVR") systems, which use speech recognition to automate customer service lines or "call centers." My experience and credentials are further detailed in the declaration that I prepared in support of ScanSoft's *Markman* brief on the '966 patent.

2. ScanSoft has now asked me to comment on U.S. Patent No. 6,594,630 ("the '630 patent"), and, in particular, how one of ordinary skill in the art (as defined in my previous declaration) would have understood that patent.

3. In forming my opinion, I have considered the following materials:

a. The '630 patent and its file history;

b. Voice Signal's May 6, 2005 *Markman* brief;

c. The Declaration of Voice Signal's expert, Charles C. Wooters;

d. Various references in the field of speech recognition as described below.

4. The '630 patent is directed to a method for controlling an electrical device through automatic speech recognition whereby syllable-length pauses between the individual words of a command phrase are utilized as part of the command phrase itself. For instance, the patent explains that a command to turn the lights on becomes "lights <pause> on" or "<pause> lights <pause> on <pause>." *See* Col. 3, lines 50-56.

5. The '630 patent alleges that the inclusion of the pauses in the command phrase results in two main benefits: First, the ability of a speech recognizer to verify the occurrence of command words in an unknown speech utterance increases with the number of syllables contained in the command because the speech recognizer is provided more information for decision making. *See* Col. 1, lines 50-54. In other words, the longer a command phrase is, the more distinct that phrase will be from normal conversation or background noise and the more accurate the speech recognizer will be in spotting it. By requiring a user to include distinct pauses between command words, the speech recognizer of the '630 patent lengthens the command, resulting in greater accuracy. Second, from a human factors perspective, requiring pauses between command words provides the speech recognizer with more information without

requiring a user to take on the burden of memorizing a longer command with additional individual words in a particular order. *See* Col. 1, lines 60-63. Thus '630 patent proposes that using pauses as part of the command allows for greater accuracy without sacrificing the user-friendliness of the speech recognition system.

6. I understand from ScanSoft's counsel and the materials that VST asserts Claims 7 and 16 of the '630 patent.

7. Claims 7 and 16 use the phrase "at least one syllable in length" to describe the minimum duration required of various speech elements (i.e. command portions and pause portions) that are processed according to the claimed methods.

8. A syllable is understood within the field of speech recognition as a perceptual unit, and not a real physical object. A linguistic definition of syllable ("… a vowel nucleus plus the optional initial and final consonants or consonant clusters") comes from the widely recognized text *Fundamentals of Speech Recognition*, by Lawrence Rabiner and Biing-Hwang Juang (Prentice Hall, 1993) on page 436 (attached as **Exhibit A**). The definition implies that vowel duration dominates the percept. A very good discussion of syllables can also be found in *Spoken Language Processing*, by Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon (Prentice Hall, 2001) in section 2.3.1 (page 51) (attached as **Exhibit B**). This text uses the example of the word "strengths" — a relatively long-duration syllable in English that can exceed half a second (more than 500 msec). *See id.* To understand the range of syllable durations, this example can be contrasted with a "schwa" – the /a/ sound at the start of the two-syllable word "about." The latter can fall under 100 msecs. As set forth by Mr. Wooters in his declaration, the average value for the duration of a syllable, however, is about 200 msecs. *See* Wooters Declaration, ¶ 4.

9. Given the widely variable lengths of syllables, from a practical perspective, the speech recognition methods recited in claims 7 and 16 of the '630 patent require a syllable-length speech element having at least a duration set above some minimum threshold. If the patent allowed for syllables of any duration, including those syllables well below the 200 msec average, the capacity of the claimed speech recognition method to accurately recognize command phrases and distinguish them from normal conversation or background noise would be severely diminished. The graph appearing at Fig. 4 of the Greenberg reference (attached to Mr. Wooters' declaration as Exhibit B), demonstrates that the average value for the duration of a syllable is roughly 200 msecs. According to that graph, there also exist syllables that range below 100 msecs or even below 50 msecs. With respect to the required pause portions, Claims 7 and 16 need a syllable-length pause of at least a minimum duration to provide the speech recognizer with more information and to clearly distinguish an intended command phrase from command words coincidentally embedded in normal conversation.

10. In light of the stated purpose of the syllable-length pauses claimed in the '630 patent and the fact that the average duration of known syllables is roughly 200 msecs, one of ordinary skill in the art would read Claims 7 and 16 to require syllable-length speech elements with a minimum duration of roughly 200 msecs.

11. Contrary to VST's assertions, one of ordinary skill in the art would not understand that the "gist of the invention" is to include syllable-length pause portions in the command phrase which are "simply the natural space between command words."[1] VST's May 6, 2005 *Markman* brief at p. 10. Rather, the '630 patent expressly defines pauses as "unnatural breaks in sound" (Col. 10, *ll.* 60-63) and teaches that the pauses are included as part of the command

[1] Indeed, even the declaration of VST's own expert, Mr. Wooters, is silent on this point and provides nothing to support this understanding of the "gist" of the invention described and claimed in the '630 patent.

phrase to provide the recognizer with more decision-making information in an effort to increase accuracy. *See* Col. 3, lines 51-60. The '630 patent plainly states: "The purpose of the speech pauses in the present invention is to make the keywords longer . . ." Col. 4, lines 5-9 (emphasis added). As the duration of the syllable-length pause decreases, the additional measure of accuracy gained by including that pause portion in the command phrase also decreases. Accordingly, one of ordinary skill in the art would understand that the syllable-length pause portions required by the asserted claims of the '630 patent must meet a minimum threshold duration and should be longer rather than shorter for the claimed speech recognition method to achieve increased accuracy.

12. Claim 7 uses the terms "spectral content" and "dynamic" to describe how pauses between spoken words are recognized and treated by the claimed speech recognition method. The recognizer of the '630 patent must correctly interpret and recognize the "silence" sounds of the pause portion between command words to accurately identify when a command phrase has been uttered. The patent describes doing this by measuring the spectral content of the spoken input over time and then comparing the pattern of change in the input to pre-stored reference templates that represent the target command phrases. When the spectral content of the required pause portion of Claim 7 is dynamic, the command phrase is not recognized and the operation of the electrical device is prevented.

13. The "spectral content" in sound is analogous to "color" in light. Sounds are vibrations in the air around us. When physical things vibrate, they don't just vibrate at a single frequency. They vibrate at many frequencies, generating harmonics and overtones—that is, a single sound is actually made up of a concurrent mixture of many different sounds at different

frequencies. The sum of all of these sounds is perceived by a human ear as the "timbre" or the "color" of the sound – its spectral content.

14. Sounds in the physical world have a very complex spectral content that can be analyzed over time. An actual complex sound wave in the real world, for example, can be said to consist of some number of simple individual sine waves (single-tone waves with no harmonics), each of which has a frequency, an amplitude (loudness), and a phase. This is also true of light waves, sound waves, surface waves on the ocean, and any other kind of wave.

15. In the field of speech recognition, it is understood that the spectral content of speech is *not* always changing and can be relatively stable over various periods of time. The "front end" of a typical speech recognizer initially converts input speech into a sequence of discrete "frames." A frame of speech represents a snapshot of how the speech looks during an analysis window of about 10 to 20 milliseconds, which is a convenient period of time for characterizing the spectral patterns that constitute the consonants and vowels of human speech. The spectral content is assumed to be changing relatively little during individual frames of speech. In addition, the state sequence models of individual phonemes contain recurrent loops back into each state because of the fact that speech is often relatively stable for sequences of multiple frames in a row, for periods up to and sometimes beyond 100 msec.

16. Thus, Mr. Wooters is misleading when he states that "the spectral content is always changing." *See* Wooters Declaration, ¶ 6. I also disagree with his statement that "In speech recognition, we use 'dynamic' to refer to changes in spectral content that are different from the change in spectral content that would be expected given the background noise." Wooters Declaration, ¶ 6. Mr. Wooters provides no reference to the formal literature of the field in support of his statements. And I am not aware of a single reference in the field that defines the

term "dynamic" as Mr. Wooters tries to, by connecting its meaning to some linked comparison to "background noise."

17. Similarly, I disagree with Mr. Wooters' unsupported assertion that one of ordinary skill in speech recognition would understand "dynamic" to mean "change in the spectral content that is different from the change in spectral content expected in the background noise." Wooters Declaration, ¶ 7. A necessary (but not sufficient) condition of Mr. Wooters' conclusion is the false assertion that "dynamic" is inherently defined in terms of "spectral content" and "background noise." It is not.

18. In speech recognition, a spectral feature that varies with time is referred to as a "dynamic" spectral feature (as opposed to a static feature that does not vary over the time analyzed). *See, e.g., Spectral Dynamics For Speech Recognition Under Adverse Conditions,* Hanson et al. in *Automatic Speech And Speaker Recognition – Advanced Topics*, by Lee et al., pp. 331-356, 1996 (attached as **Exhibit C**). This use of the term "dynamic" to refer to a spectral feature that varies over time is widely recognized throughout the field of speech recognition and, if need be, I could provide citations to dozens of peer-reviewed articles that do so.[2] Simply put, within the context of Claim 7, stable spectral content indicates a pause or silence, while dynamic spectral content indicates the absence of a pause. This understanding of the term "dynamic" within the field is not related to any consideration of "background noise." Mr. Wooters is wrong to suggest otherwise.

19. The preambles of Claims 7 and 16 use the term "activate." Contrary to Mr. Wooter's Declaration, that term has no established, formal definition within the field of speech recognition. *See* Wooters Declaration, ¶ 3. To explore the subject, I researched various well-

respected texts (e.g. Rabiner and Juang, *Fundamentals of Speech Recognition*; Huang, Acero, and Hon, *Spoken Language Processing*) on speech recognition as well as related texts on user interface design and command-and-control systems. None of these texts mentioned the word "activate" in their index or table of contents, nor was I able to find specific discussions within any text that defined the phrase. Accordingly, I conclude that one of ordinary skill in the art would have understood the term "activate" to have its ordinary English usage as set forth in a dictionary.

[2] This discussion of the term "dynamic" is with reference to spectral activity in speech, as it is used in claim 7 of the '630 patent. In other contexts, it may convey different meanings; for example, in the phrase "Dynamic Time Warping."

I DECLARE UNDER PENALTY OF PERJURY THAT THE FOREGOING IS TRUE AND CORRECT. EXECUTED ON JUNE 3, 2005.

/s/ Bruce Balentine
BRUCE BALENTINE

02639/00509 387606.1

I DECLARE UNDER PENALTY OF PERJURY THAT THE FOREGOING IS TRUE AND CORRECT. EXECUTED ON JUNE 3, 2005.

/s/
BRUCE BALENTINE

02639/00509 387606.1